PHOTO MAKER:

通过堆叠 ID 嵌入实现逼真人像照片定制

Zhen Li[1,2 *] Mingdeng Cao[2,3 *] Xintao Wang[2 †] Zhongang Qi[2] Ming-Ming Cheng[1 †] Ying Shan[2]

[1]VCIP, CS, Nankai University [2]ARC Lab, Tencent PCG [3]The University of Tokyo

https://photo-maker.github.io/

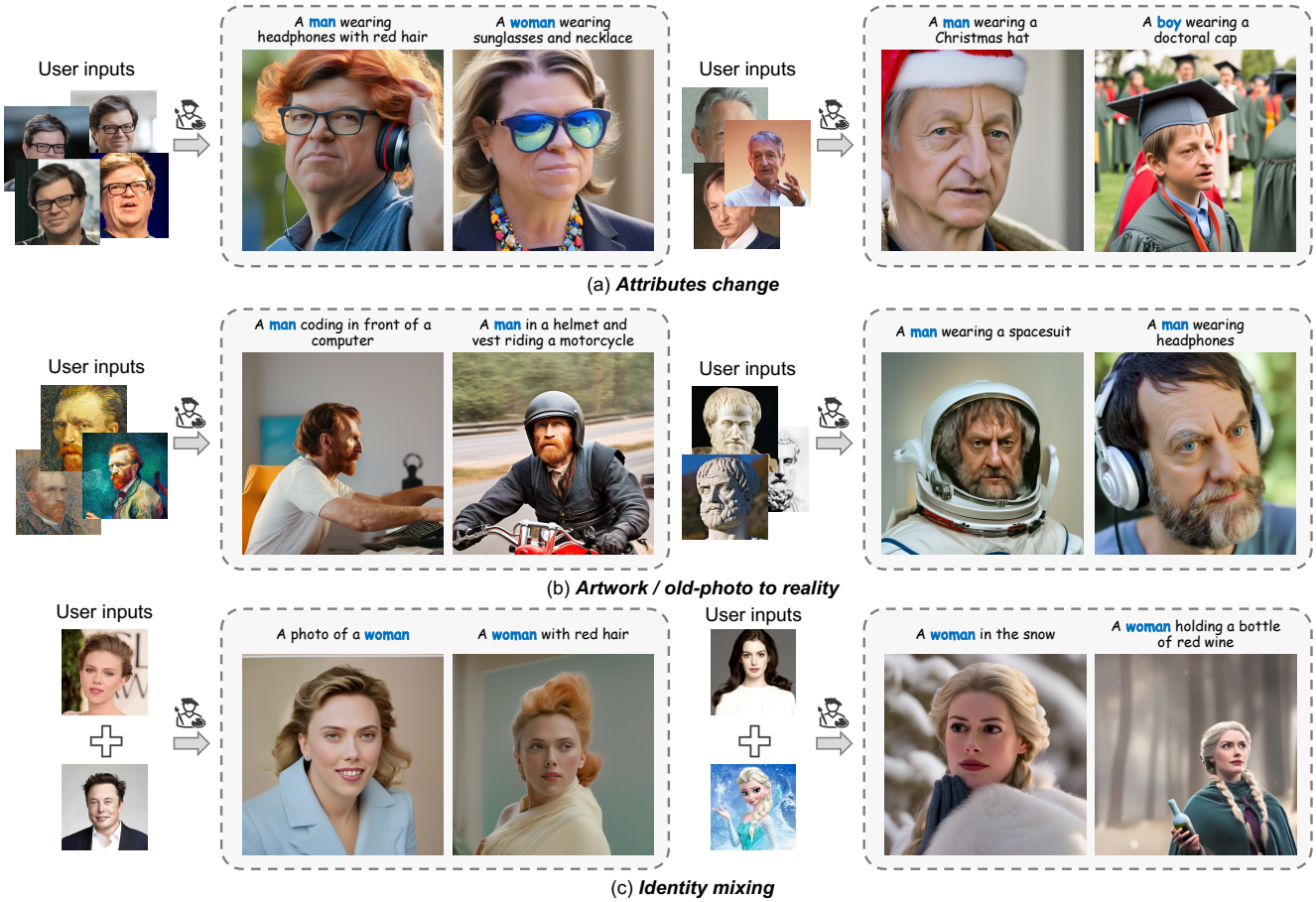


图 1. 给定一些带有 ID 的图像，所提出的 PhotoMaker 能够**在单次前向传播**中，根据文本提示生成多样化的个性化 ID 图像。我们的方法不仅能够很好地保留输入图像集中的 ID 信息，还能生成逼真的人像照片。PhotoMaker 还支持多种有趣的应用，例如：(a) 更改属性，(b) 将艺术作品或老照片中的人物带入现实，或 (c) 执行身份混合。

Abstract

近年来，文本生成图像技术取得了显著进展，能够根据给定的文本提示合成逼真的人像照片。然而，现有的个性化生成方法往往无法同时满足高效率、卓越的身份（ID）保真度和灵活的文本控制能力。在本工作中，我们提出了 PhotoMaker，一种高效的个性化文本生成图像方法。通过将任意数量的带有 ID 的图像编码为堆叠的 ID 嵌入，我们能够保留 ID 信息。这种嵌入作为统一的 ID 表示，不仅能全面地封装相同输入 ID

* Interns in ARC Lab, Tencent PCG † Corresponding authors

的特征，还能兼容不同 ID 的特征，便于后续的整合。这为更多引人入胜且具有实际价值的应用提供了可能。此外，为了促进我们 PhotoMaker 的训练，我们提出了一种面向 ID 的数据构建流程，用于组装训练数据。在这一流程构建的数据集的支持下，我们的 PhotoMaker 表现出了比基于测试时微调的方法更强的 ID 保真度，同时提供了显著的速度提升、高质量的生成效果、强大的泛化能力以及广泛的应用潜力。

1. Introduction

与人类相关的定制化图像生成 [26, 33, 50] 受到了广泛关注，催生了许多应用，如个性化人像照片 [35]、图像动画 [67] 和虚拟试穿 [59]。早期的方法 [38, 40] 由于生成模型（i.e., GANs [16, 27]）能力的限制，只能定制生成面部区域，导致多样性、场景丰富性和可控性较低。得益于大规模文本-图像对训练数据集 [55]、更大规模的生成模型 [43, 52] 以及能够提供更强语义嵌入的文本/视觉编码器 [44, 45]，基于扩散的文本生成图像模型在最近不断演进。这一演进使其能够生成愈加逼真的面部细节和丰富的场景。由于文本提示和结构引导的存在，生成的可控性也得到了极大提升 [39, 65]。

与此同时，在强大的扩散文本生成图像模型的滋养下，许多基于扩散的定制化生成算法 [14, 49] 应运而生，以满足用户对高质量定制化结果的需求。在商业和社区应用中，最广泛使用的是基于 DreamBooth 的方法 [49, 51]。这些应用需要相同身份（ID）的几十张图像来微调模型参数。虽然生成的结果具有高 ID 保真度，但存在两个明显的缺点：一是每次用于微调的定制数据需要人工收集，因此非常耗时费力；二是定制每个 ID 需要 10-30 分钟，消耗大量计算资源，特别是当生成模型变得更大时。因此，为了简化并加速定制生成过程，受现有以人为中心的数据集 [27, 34] 的驱动，近年来的研究通过训练视觉编码器 [8, 62] 或超网络 [2, 50] 将输入 ID 图像表示为模型的嵌入或 LoRA [22] 权重。在训练之后，用户只需提供一个 ID 图像，即可通过少量的微调步骤甚至无需微调实现个性化生成。然而，这些方法定制的结果无法像 DreamBooth 那样同时具备 ID 保真度和生成多样性（见 Fig. 3）。这是因为：1）在训练过程中，目标图像和输入的 ID 图像从同一图像采样，训练好的模型容易记住与 ID 无关的特征（如表情和视角），导致可编辑性差；2）仅依赖单个 ID 图像进行定制，使模型难以从其内部知识中辨别要生成的 ID 特征，导致 ID 保真度不佳。

基于以上两点，并受到 DreamBooth 成功的启发，我们在本文中旨在：1）确保作为控制的 ID 图像和目标图像在视角、面部表情和配饰上呈现出变化，使模型不记住与 ID 无关的信息；2）在训练过程中为模型提供同一 ID 的多张不同图像，以更全面准确地表示定制化 ID 的特征。

因此，我们提出了一种简单但有效的前向定制化人类图像生成框架，称之为 PhotoMaker，能够接收多个输入 ID 图像。为更好地表示每张输入图像的 ID 信息，我们在语义层面堆叠多个输入 ID 图像的编码，构建堆叠 ID 嵌入。该嵌入可被视为要生成的 ID 的统一表示，其中每个子部分对应一个输入 ID 图像。为了更好地将该 ID 表示与文本嵌入整合到网络中，我们用堆叠 ID 嵌入替换文本嵌入中的类别词（例如，男人和女人）。这样，结果嵌入同时表示要定制的 ID 和要生成的上下文信息。通过这种设计，在不向网络中添加额外模块的情况下，生成模型的交叉注意力层可以自适应地整合堆叠 ID 嵌入中包含的 ID 信息。

同时，堆叠 ID 嵌入允许我们在推理时接收任意数量的 ID 图像作为输入，并保持其他免微调方法 [56, 62] 相似的生成效率。具体而言，当接收四张 ID 图像时，我们的方法生成一张定制化人像约需 10 秒，比 DreamBooth 快约 130 倍[1]。此外，由于我们的堆叠 ID 嵌入可以更全面、准确地表示定制化 ID，与最先进的免微调方法相比，我们的方法提供了更好的 ID 保真度和生成多样性。与先前的方法相比，我们的框架在可控性方面也有了很大提升。不仅可以执行常见的重上下文化，还可以更改输入人像的属性（如，配饰和表情）、生成与输入 ID 完全不同视角的人像，甚至可以修改输入 ID 的性别和年龄（见 Fig. 1）。

值得注意的是，我们的 PhotoMaker 还为用户生成定制化人像图像提供了很多可能性。具体而言，虽然构建堆叠 ID 嵌入的图像在训练过程中来自同一 ID，但在推理时我们可以使用不同的 ID 图像来形成堆叠 ID 嵌入，以合并并创建一个新的定制 ID。合并后的新 ID 可以保留不同输入 ID 的特征。例如，我们可以生成具有 Elon Musk 特征的 Scarlett Johansson，或将人物与知名 IP 角色混合（见 Fig. 1(c)）。同时，合并比例可

[1]在一张 NVIDIA Tesla V100 上测试

以通过提示加权 [18, 23] 或改变输入图像集中不同 ID 图像的比例来简单调整，展示了我们框架的灵活性。

我们的 PhotoMaker 在训练过程中需要同时输入多张具有相同 ID 的图像，因此需要 ID 导向的人类数据集的支持。然而，现有的数据集要么未按 ID 分类 [27, 33, 55, 68]，要么只关注人脸而不包括其他上下文信息 [34, 40, 60]。因此，我们设计了一条自动化流程来构建一个 ID 相关的数据集，以便于训练我们的 PhotoMaker 。通过该流程，我们可以构建包含大量 ID 的数据集，每个 ID 具有多张图像，包含丰富的视角、属性和场景。同时，在此流程中，我们可以为每张图像自动生成描述，并标注相应的类别词 [49]，以更好地适应我们框架的训练需求。

2. Related work

文本生成图像的扩散模型. 扩散模型 [20, 58] 在文本生图方面取得了显著进展 [28, 46, 48, 52]，近年来吸引了广泛关注。这些模型的优异性能归因于高质量的大规模文本-图像数据集 [6, 54, 55]、基础模型的持续升级 [7, 42]、条件编码器 [24, 44, 45]，以及可控性的提升 [32, 39, 63, 65]。基于这些进展，Podell et al. [43] 开发了目前最强大的开源生成模型 SDXL。鉴于其在生成人像方面的出色表现，我们基于此模型构建了 PhotoMaker。然而，我们的方法也可以扩展至其他文生图模型。

扩散模型中的个性化. 由于扩散模型强大的生成能力，越来越多的研究人员尝试基于扩散模型探索个性化生成。目前，主流的个性化生成方法主要可以分为两类。一类依赖于测试阶段的额外优化，例如 DreamBooth [49] 和 Textual Inversion [14]。由于这两项开创性工作都需要大量时间进行微调，一些研究尝试通过减少调优所需参数数量 [17, 30, 51, 64] 或通过大规模数据集的预训练 [15, 50] 来加速个性化定制的过程。尽管取得了进展，这些方法仍需要对每个新概念对预训练模型进行微调，这一过程耗时，并限制了其应用。最近，一些研究 [9, 10, 25, 36, 37, 56, 61] 尝试使用单张图像在一次前向传播中进行个性化生成，显著加快了个性化进程。这些方法要么利用个性化数据集 [9, 57] 进行训练 ，要么将需要定制的图像编码到语义空间 [8, 25, 37, 56, 61, 62]。我们的方法结合了上述两种技术。具体来说，我们不仅依赖于构建一个面向 ID 的个性化数据集，还依赖于获取代表人物 ID 的语义嵌入。与以往基于嵌入的方法不同，我们的 PhotoMaker 通过从多张 ID 图像中提取堆叠 ID 嵌入，不仅提供了更好的 ID 表示，同时保持了与之前基于嵌入的方法相同的高效率。

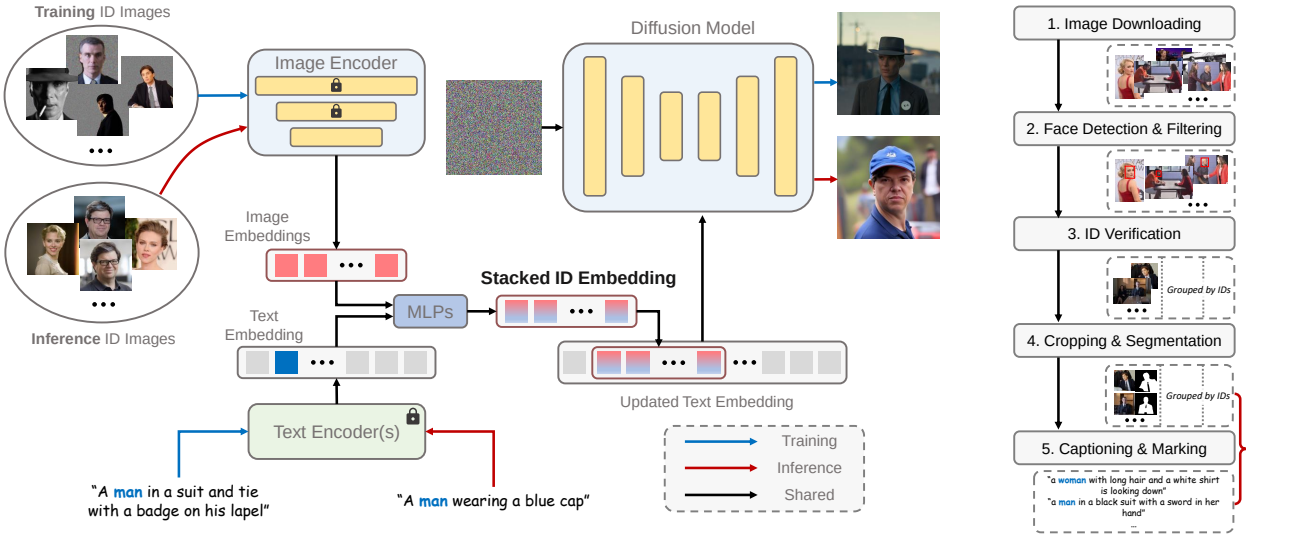
3. Method

3.1. Overview

给定需要定制的 ID 图像，PhotoMaker 旨在生成一张新的逼真人像图像，该图像不仅能够保留输入人像的身份的特征，还能够根据文本提示的控制改变生成图像内容或属性。尽管我们像 DreamBooth 一样输入多张人像进行定制，但我们仍然保持与其他免微调方法相同的高效性，通过单次前向传播完成定制，同时确保出色的 ID 保真度和文本可编辑性。此外，我们还可以混合多个输入 ID，生成的图像能够很好地保留不同 ID 的特征，从而为更多应用提供可能性。上述能力主要得益于我们提出的简单而有效的堆叠 ID 嵌入，它能够为输入 ID 提供统一的表示。为了便于训练我们的 PhotoMaker，我们设计了一条数据构建流程，建立一个按 ID 分类的以人为中心的数据集。Fig. 2(a) 展示了所提出的 PhotoMaker 概览，Fig. 2(b) 展示了我们的数据构建流程。

3.2. Stacked ID Embedding

编码器. 跟随近期的研究工作 [25, 56, 61]，我们使用 CLIP [44] 图像编码器 $\mathcal{E}img$ 来提取图像嵌入，为了与扩散模型中的原始文本表示空间对齐。在将每个输入图像输入图像编码器之前，我们将特定 ID 的身体部分之外的图像区域填充为随机噪声，以消除其他 ID 和背景的影响。由于用于训练原始 CLIP 图像编码器的数据主要由自然图像组成，为了更好地使模型从被掩码的图像中提取与 ID 相关的嵌入，我们在训练我们的 PhotoMaker 时对图像编码器中的部分 transformer 层进行了微调。我们还引入了额外的可学习的 project 层，将从图像编码器获得的嵌入映射到与文本嵌入相同的维度中。设 $\{X^i \mid i = 1 \dots N\}$ 表示从用户获取的 N 个输入 ID 图像，获得提取的嵌入 $\{e^i \in \mathbb{R}^D \mid i = 1 \dots N\}$，其中 D 表示 project 后的维度。每个嵌入对应于输入图像的 ID 信息。对于给定的文本提示 T，我们使用预训练的 CLIP 文本编码器 $\mathcal{E}text$ 提取文本嵌入 $t \in \mathbb{R}^{L \times D}$，其中 L 表示嵌入的长度。



(a) (b)

图 2. **提出的** (a) PhotoMaker **和** (b) **面向** ID **的数据构建流程概览**。对于提出的 PhotoMaker，我们首先分别从文本编码器和图像编码器获取文本嵌入和图像嵌入。然后，我们通过合并对应的类别嵌入（例如，man 和 woman）与每个图像嵌入来提取融合嵌入。接下来，我们沿着长度维度将所有融合嵌入拼接起来，形成堆叠 ID 嵌入。最后，我们将堆叠 ID 嵌入输入到所有交叉注意力层中，以自适应地将 ID 信息融入扩散模型中。请注意，尽管我们在训练过程中使用了相同 ID 的带有掩码背景的图像，但在推理阶段，我们可以直接输入不同 ID 的图像，并在不产生背景失真的情况下创建新的 ID。

堆叠. 近期的研究工作 [14, 49, 62] 表明，在文生图模型中，个性化角色 ID 信息可以被一些独特的标记表示。我们的方法也有类似的设计，以更好地表示输入人像的 ID 信息。具体来说，我们在输入的文本中标记对应的类别词（例如，man 和 woman）（参见 Sec. 3.3）。然后，我们提取文本嵌入中类别词对应位置的特征向量。这个特征向量将与每个图像嵌入 e^i 融合。我们使用两个 MLP 层来执行这种融合操作。融合后的嵌入可以表示为 $\{\hat{e}^i \in \mathbb{R}^D \mid i = 1 \dots N\}$。通过结合类别词的特征向量，这个嵌入可以更全面地表示当前输入的 ID 图像。此外，在推理阶段，这种融合操作还为定制化生成过程提供了更强的语义可控性。例如，我们可以通过简单地替换类别词来定制人物 ID 的年龄和性别（参见 Sec. 4.2）。

在获得融合后的嵌入后，我们沿着长度维度将它们连接起来，形成堆叠的 ID 嵌入：

$$s^* = \text{Concat}([\hat{e}^1, \dots, \hat{e}^N]) \quad s^* \in \mathbb{R}^{N \times D}. \tag{1}$$

这个堆叠的 ID 嵌入可以作为多个 ID 图像的统一表示，同时保留每个输入 ID 图像的原始表示。它可以接受任意数量的 ID 图像编码嵌入，因此其长度 N 是可变的。与基于 DreamBooth 的方法 [49, 51] 不同，这些方法将多个图像输入模型进行个性化定制微调，而我们的方法本质上是同时将多个嵌入输入模型。在将相同 ID 的多个图像打包成一个 batch 作为图像编码器的输入之后，堆叠的 ID 嵌入可以通过一次前向传递获得，这相比于基于微调的方法显著提高了效率。同时，与其他基于嵌入的方法 [61, 62] 相比，这种统一的表示能够同时保持较好的 ID 保真度和文本可控性，因为它包含了更全面的 ID 信息。此外，值得注意的是，尽管我们在训练期间仅使用了相同 ID 的多张图像来构建这个堆叠 ID 嵌入，但在推理阶段我们可以使用来自不同 ID 的图像来构建它。这种灵活性为许多有趣的应用打开了可能性。例如，我们可以混合现实中存在的两个人，或者将一个人和一个著名的角色 IP 混合（参见 Sec. 4.2）。

合并. 我们使用扩散模型中固有的交叉注意力机制来自适应地融入堆叠 ID 嵌入中包含的 ID 信息。我们首先将原始文本嵌入 t 中对应于类别词的位置的特征向量替换为堆叠 ID 嵌入 s^*，从而得到更新后的文本嵌入 $t^* \in \mathbb{R}^{(L+N-1) \times D}$。然后，交叉注意力操作可以表示为：

$$\begin{cases} \text{Q} = \text{W}_Q \cdot \phi(z_t); \ \text{K} = \text{W}_K \cdot t^*; \ \text{V} = \text{W}_V \cdot t^* \\ \text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}(\frac{\text{QK}^T}{\sqrt{d}}) \cdot \text{V}, \end{cases} \tag{2}$$

其中，$\phi(\cdot)$ 是一个可以通过 UNet 去噪器从输入潜在空

间编码得到的嵌入。W_Q、W_K 和 W_V 是映射矩阵。此外，我们可以通过文本加权 [18, 23] 调整每个输入 ID 图像在生成新定制 ID 时的参与程度，这展示了我们 PhotoMaker 的*灵活性*。近期的研究工作 [30, 51] 发现，通过简单地调整注意力层的权重，可以实现良好的 ID 定制化性能。为了让原始扩散模型更好地感知堆叠 ID 嵌入中包含的 ID 信息，我们还额外训练了注意力层中矩阵的 LoRA [22, 51] 残差。

3.3. ID-Oriented Human Data Construction

由于我们的 PhotoMaker 在训练过程中需要采样相同 ID 的多个图像来构建堆叠 ID 嵌入，因此我们需要使用按 ID 分类的数据集来驱动 PhotoMaker 的训练过程。然而，现有的人类数据集要么没有标注 ID 信息 [27, 33, 55, 68]，要么它们所包含的场景的丰富性非常有限 [34, 40, 60]（即，它们仅关注面部区域）。因此，在本节中，我们将介绍一个构建以人为中心的文本-图像数据集的流程，该数据集按不同 ID 分类。Fig. 2(b) 展示了所提出的流程。通过这个流程，我们可以收集一个以 ID 为导向的数据集，其中包含大量的 ID，每个 ID 有多张图像，包括不同的表情、属性、场景等。这个数据集不仅便于我们的 PhotoMaker 的训练过程，还可能启发未来潜在的基于 ID 的研究。数据集的统计信息显示在附录中。

图像下载. 我们首先列出了一个名人名单，这些名单可以从 VoxCeleb1 和 VGGFace2 [4] 中获取。我们根据名单中的名字在搜索引擎中进行搜索并爬取数据。每个名字大约下载了 100 张图像。为了生成更高质量的肖像图像 [43]，我们在下载过程中过滤了分辨率最短边小于 512 的图像。

人了检测与过滤. 我们首先使用 RetinaNet [13] 来检测人脸边界框，并过滤掉尺寸较小的检测结果（小于 256 × 256）。如果图像中没有任何符合要求的边界框，则该图像将被过滤掉。然后，我们对剩余的图像进行 ID 验证。

ID **验证**. 由于一张图像可能包含多个面部，我们首先需要确定哪个面部属于当前的身份组。具体来说，我们将当前身份组中检测框内的所有面部区域送入 ArcFace [12] 提取身份嵌入，并计算每对面部的 L2 相似度。我们将每个身份嵌入与所有其他嵌入计算的相似度求和，以获得每个边界框的得分。对于包含多个面部的图像，我们选择得分最高的边界框。边界框选择后，我们重新计算每个剩余框的总得分。我们计算每个 ID 组的总得分的标准差 δ。我们通过经验使用 8δ 作为阈值来过滤掉 ID 不一致的图像。

裁剪与分割. 我们首先基于检测到的面部区域裁剪出一个较大的方形框，同时确保裁剪后的面部区域能够占据图像的 10% 以上。由于我们需要在将输入 ID 图像送入图像编码器之前去除与背景和 ID 无关的部分，因此需要为指定 ID 生成掩码。具体来说，我们使用 Mask2Former [11] 对 “person” 类别进行全景分割。我们保留与面部边界框重叠度最高的掩码。此外，我们选择丢弃没有检测到掩码的图像，以及边界框与掩码区域之间没有重叠的图像。

文本与标注. 我们使用 BLIP2 [31] 为每张裁剪后的图像生成文本。由于我们需要标记类别词（例如，man、woman 和 boy）以便于文本和图像嵌入的融合，我们通过 BLIP2 的随机模式重新生成不包含任何类别词的文本，直到出现类别词为止。在获得文本后，我们将文本中的类别词单数化，以便专注于单一的 ID。接下来，我们需要标记与当前 ID 对应的类别词位置。对于仅包含一个类别词的文本，可以直接进行标注。对于包含多个类别词的文本，我们统计每个身份组中文本中包含的类别词。出现次数最多的类别词将作为当前身份组的类别词。然后，我们使用每个身份组的类别词来匹配并标记该身份组中的每个文本。对于不包含与对应身份组匹配的类别词的文本，我们使用依赖句法分析模型 [21] 根据不同的类别词对文本进行分割。我们计算分割后的子文本与图像中特定 ID 区域之间的 CLIP 分数 [44]。此外，我们通过 SentenceFormer [47] 计算当前分割部分的类别词与当前身份组的类别词之间的标签相似度。我们选择标记 CLIP 分数和标签相似度的乘积最大的类别词。

4. Experiments

4.1. Setup

配置细节. 为了生成更具照片真实感的人物肖像，我们采用了 SDXL 模型 [43] stable-diffusion-xl-base-1.0 作为我们的文生图模型。相应地，训练数据的分辨率被调整为 1024×1024。我们使用 CLIP ViT-L/14 [44] 和一个额外的 project 层来获取初始图像嵌入 e^i。对于

文本嵌入，我们保留 SDXL 中的原始两个文本编码器进行提取。整个框架使用 Adam 优化器 [29] 在 8 个 NVIDIA A100 GPU 上训练两周，批大小为 48。我们将 LoRA 权重的学习率设置为 $1e-4$，其他可训练模块的学习率设置为 $1e-5$。在训练过程中，我们随机采样 1-4 张与当前目标 ID 图像相同 ID 的图像来形成堆叠 ID 嵌入。此外，为了通过无分类器引导提高生成性能，我们有 10% 的概率使用空文本嵌入替换原始更新的文本嵌入 t^*。我们还使用带有 50% 概率的掩码扩散损失 [3]，鼓励模型生成更真实的 ID 相关区域。在推理阶段，我们使用 delayed subject conditioning [62] 来解决文本和 ID 条件之间的冲突。我们使用 50 步的 DDIM 采样器 [58]，并将无分类器引导的比例设置为 5。

评估指标. 跟随 DreamBooth [49]，我们使用 DINO [5] 和 CLIP-I [14] 指标来衡量 ID 保真度，并使用 CLIP-T [44] 指标来衡量提示保真度。为了更全面的评估，我们还通过检测和裁剪生成图像与真实图像中相同 ID 的面部区域来计算面部相似度。我们使用 RetinaFace [13] 作为检测模型，面部嵌入通过 FaceNet [53] 提取。为了评估生成质量，我们采用 FID 指标 [19, 41]。重要的是，由于大多数基于嵌入的方法倾向于将面部姿势和表情融入表示中，生成的图像通常缺乏面部区域的变化。因此，我们提出了一种指标，称为*面部多样性*，用于衡量生成面部区域的多样性。具体来说，我们首先检测并裁剪每张生成图像中的面部区域。接下来，我们计算所有生成图像中每对面部区域之间的 LPIPS [66] 分数，并取平均值。该值越大，生成的面部区域多样性越高。

Evaluation dataset. 我们的评估数据集包括 25 个 ID，其中包含 9 个来自 Mystyle [40] 的 ID，以及我们自己收集的另外 16 个 ID。请注意，这些 ID 在训练集中没有出现，旨在评估模型的泛化能力。为了进行更全面的评估，我们还准备了 40 个提示，涵盖了各种表情、属性、装饰、动作和背景。对于每个 ID 的每个提示，我们生成 4 张图像进行评估。更多细节列在附录中。

4.2. Applications

在本节中，我们将详细阐述我们的 PhotoMaker 可以赋能的应用。对于每个应用，我们选择最适合该设置的比较方法。比较方法将从 DreamBooth [49]、Textual Inversion [14]、FastComposer [62] 和 IPAdapter [63] 中选择。我们优先使用每种方法提供的官方模型。对于 DreamBooth 和 IPAdapter，我们使用它们的 SDXL 版本以进行公平比较。对于所有应用，我们选择了四张输入 ID 图像来形成我们 PhotoMaker 中的堆叠 ID 嵌入。我们还公平地使用四张图像来训练需要测试时优化的方法。我们在附录中为每个应用提供了更多样本。

重新语境化我们首先展示了通过简单的语境变化（例如修改发色和衣物，或基于基本提示控制生成背景）得到的结果。由于所有方法都可以适应此应用，我们对生成结果进行了定量和定性的比较（见 Tab. 1和 Fig. 3）。结果表明，我们的方法能够很好地满足生成高质量图像的能力，同时确保较高的 ID 保真度（具有最高的 CLIP-T 和 DINO 分数，并且是第二高的面部相似度）。与大多数方法相比，我们的方法生成的图像质量更高，生成的面部区域表现出更大的多样性。与此同时，我们的方法能够保持与基于嵌入的方法一致的高效率。为了更全面的比较，我们在附录的 Sec. B中展示了用户研究结果。

将艺术作品/老照片中的人物带回现实通过将艺术画作、雕塑或旧照片作为输入，我们的 PhotoMaker 可以将来自上个世纪甚至古代的人物带到现代，为他们“拍照”。Fig. 4(a) 展示了结果。与我们的方法相比，DreamBooth 和 SDXL 在生成未出现在真实照片中的逼真人物图像时存在困难。此外，由于 DreamBooth 过度依赖定制图像的质量和分辨率，当使用旧照片进行定制生成时，DreamBooth 很难生成高质量的结果。

改变年龄或性别通过简单地替换类别词（例如 man 和 woman），我们的方法可以实现性别和年龄的变化。Fig. 4(b) 展示了结果。尽管 SDXL 和 DreamBooth 在提示工程后也能实现相应的效果，但由于堆叠 ID 嵌入的作用，我们的方法能够更容易地捕捉人物的特征信息。因此，我们的结果显示出更高的 ID 保真度。

身份混合如果用户提供不同 ID 的图像作为输入，我们的 PhotoMaker 可以很好地整合不同 ID 的特征，形成一个新的 ID。从 Fig. 5中可以看出，DreamBooth 和 SDXL 都无法实现身份混合。相比之下，我们的方法能够在生成的新 ID 上很好地保留不同 ID 的特征，无论输入的是动漫 IP 还是真人人物，也不受性别的限制。此外，我们可以通过控制对应 ID 的输入数量或提示加权来控制新生成 ID 中该 ID 的比例。我们在附录中的

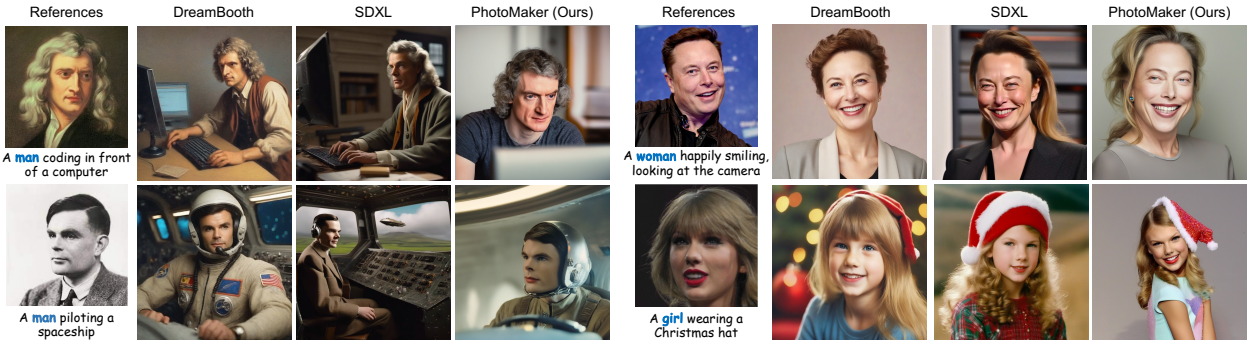
图 3. **定性比较**。我们将我们的方法与 DreamBooth [49]、Textual Inversion [14]、FastComposer [62] 和 IPAdapter [63] 进行比较，使用五个不同的身份和相应的提示。我们观察到，我们的方法通常能够实现高质量的生成，良好的可编辑性和强大的身份保真度。

	CLIP-T↑ (%)	CLIP-I↑ (%)	DINO↑ (%)	Face Sim.↑ (%)	Face Div.↑ (%)	FID↓	Speed↓ (s)
DreamBooth [49]	29.8	62.8	39.8	49.8	49.1	374.5	1284
Textual Inversion [14]	24.0	70.9	39.3	54.3	59.3	363.5	2400
FastComposer [62]	<u>28.7</u>	66.8	40.2	61.0	45.4	375.1	8
IPAdapter [63]	25.1	<u>71.2</u>	<u>46.2</u>	67.1	52.4	375.2	12
PhotoMaker (Ours)	26.1	73.6	51.5	<u>61.8</u>	<u>57.7</u>	<u>370.3</u>	10

表 1. **定量比较**。用于基准测试的指标包括保留 ID 信息的能力（即，CLIP-I、DINO 和面部相似度）、文本一致性（即，CLIP-T）、生成面部的多样性（即，面部多样性）和生成质量（即，FID）。此外，我们定义个性化速度为在输入 ID 控制后获得最终个性化图像所需的时间。我们在单个 NVIDIA Tesla V100 GPU 上测量个性化时间。最佳结果以**粗体**显示，第二好结果以<u>下划线</u>显示。

Fig. 10-11展示了这一能力。

风格化在 Fig. 6中，我们展示了我们方法的风格化能力。可以看出，在生成的图像中，我们的 PhotoMaker 不仅保持了良好的 ID 保真度，还有效地展现了输入提



(a) (b)

图 4. **在（a）艺术作品和旧照片，以及（b）更改年龄或性别上的应用**。我们能够将过去的人物带回现实生活或更改输入 ID 的年龄和性别。对于第一个应用，我们为 DreamBooth 和 SDXL 准备了文本模板 A photo of , photo-realistic。相应地，我们将原始文本中的类别词更改为名人名称。对于第二个应用，我们将类别词替换为 <class word> <name>, (at the age of 12)

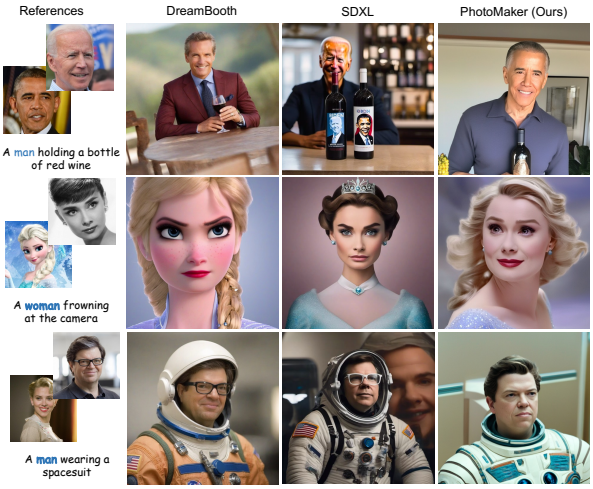


图 5. **身份混合**。我们能够生成具有新 ID 的图像，同时保留输入身份特征。我们为 SDXL 准备了提示模板 <original prompt>, with a face blended with name:A and name:B

图 6. PhotoMaker **的风格化结果**。符号 <class> 表示它将根据情况被替换为 man 或 woman。

示的风格信息。这揭示了我们的方法在推动更多应用方面的潜力。附录中的 Fig. 12展示了更多结果。

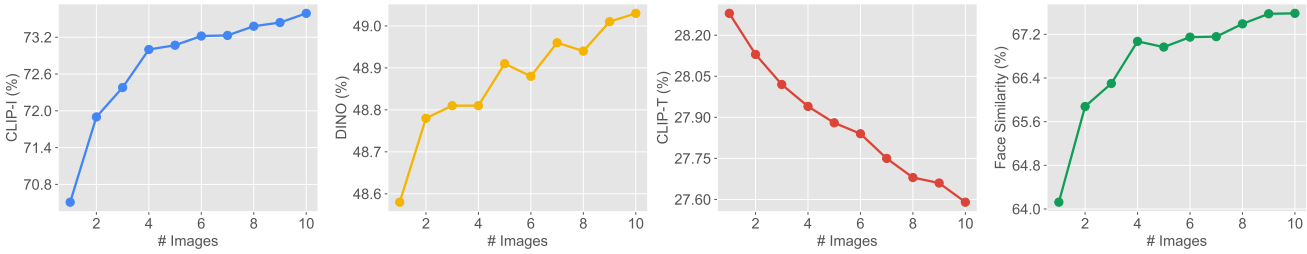
4.3. Ablation study

为了对每个变体进行消融研究，我们将总训练迭代次数缩短了八倍。

输入 ID 图像数量的影响. 我们探索了通过输入不同数量的 ID 图像来形成所提出的堆叠 ID 嵌入的影响。在 Fig. 7中，我们可视化了这一影响在不同指标上的表现。我们得出结论，使用更多的图像来形成堆叠 ID 嵌入可以提高与 ID 保真度相关的指标。特别是在输入图像数量从一张增加到两张时，这种提升尤为明显。随着输入 ID 图像数量的增加，ID 相关指标的值增长速率显著减缓。此外，我们还观察到 CLIP-T 指标呈线性下降。这表明文本可控性和 ID 保真度之间可能存在权衡。从 Fig. 8中，我们可以看到，增加输入图像的数量增强了 ID 的相似性。因此，更多的 ID 图像用于形成堆叠 ID 嵌入可以帮助模型感知更全面的 ID 信息，从而更准确地表示 ID 以生成图像。此外，正如 Dwayne Johnson 示例所示，性别编辑能力有所下降，模型更容易生成原始 ID 性别的图像。

组合多个嵌入的选择我们探索了三种构建 ID 嵌入的方法，包括对图像嵌入进行平均、通过线性层自适应投影嵌入，以及我们的堆叠方式。从 Tab. 2a中，我们可以看到，堆叠方式在确保生成面部的多样性的同时，具有最高的 ID 保真度，证明了其有效性。此外，这种方式比其他方法提供了更大的灵活性，包括接受任意数量的图像以及更好地控制不同 ID 的混合过程。

训练过程中多个嵌入的好处我们探索了另外两种训练数据采样策略，以证明在训练过程中输入具有变化的多张图像是必要的。第一种策略是仅选择一张图像，这



(a) (b) (c) (d)

图 7. **输入 ID 图像数量对** (a) CLIP-I，(b) DINO，(c) CLIP-T **和** (d) **面部相似度的影响**。

	CLIP-T↑	DINO↑	Face Sim.↑	Face Div.↑
Average	28.7	47.0	48.8	56.3
Linear	28.6	47.3	48.1	54.6
Stacked	28.0	49.5	53.6	55.0

(a) **嵌入组合选择**.

	CLIP-T↑	DINO↑	Face Sim.↑	Face Div.↑
Single embed	27.9	50.3	50.5	56.1
Single image	27.3	50.3	60.4	51.7
Ours	28.0	49.5	53.6	55.0

(b) **训练数据采样策略**.

表 2. **提出的** PhotoMaker **的消融研究**。最佳结果用**粗体**标记。

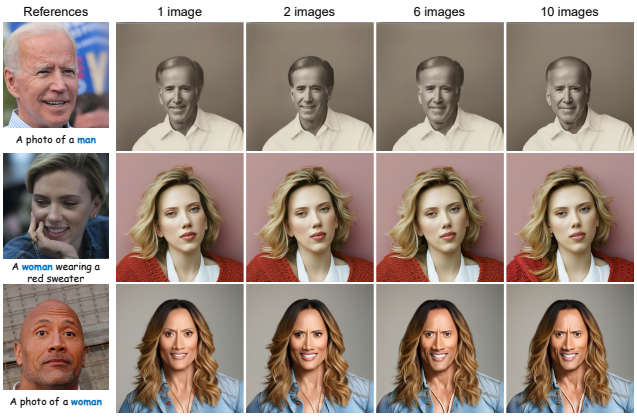


图 8. **输入图像数量变化对生成结果的影响**。可以观察到，随着输入图像数量的增加，ID 的保真度也随之提高。

张图像可以与目标图像不同，用来形成 ID 嵌入（见 Tab. 2b中的 “single embed”）。我们的方法在 ID 保真度上具有优势。第二种采样策略是将目标图像作为输入 ID 图像（模拟大多数基于嵌入的方法的训练方式）。我们基于这张图像使用不同的数据增强方法生成多张图像，并提取相应的多个嵌入。在 Tab. 2b中，由于模型容易记住输入图像的其他无关特征，生成的面部区域缺乏足够的变化（多样性低）。

5. Conclusion

我们提出了 PhotoMaker，一种高效的个性化文本到图像生成方法，专注于生成真实的人物照片。我们的方法利用了一种简单而有效的表示方式——堆叠 ID 嵌入，以更好地保留 ID 信息。实验结果表明，与其他方法相比，我们的 PhotoMaker 能够同时满足高质量和多样化的生成能力，具有良好的可编辑性、高效的推理速度和强大的 ID 保真度。此外，我们还发现，我们的方法可以赋能许多以前的方法难以实现的有趣应用，如改变年龄或性别、将人物从旧照片或艺术作品带回现实以及身份混合。

参考文献

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. TOG, 2021. 4

[2] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. TOG, 2023. 2

[3] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In SIGGRAPH Asia, 2023. 6

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and

Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In FG, 2018. 5

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In ICCV, 2021. 6, 1

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In CVPR, 2021. 3

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426, 2023. 3

[8] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. arXiv preprint arXiv:2309.05793, 2023. 2, 3

[9] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. arXiv preprint arXiv:2304.00186, 2023. 3

[10] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481, 2023. 3

[11] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In CVPR, 2022. 5

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, 2019. 5, 1

[13] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In CVPR, 2020. 5, 6

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In ICLR, 2023. 2, 3, 4, 6, 7, 1

[15] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. arXiv preprint arXiv:2302.12228, 2023. 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. ACM Communications, 2020. 2

[17] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In ICCV, 2023. 3

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In ICLR, 2023. 3, 5, 2

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In NeurIPS, 2017. 6

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020. 3

[21] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python, 2020. 5

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In ICLR, 2022. 2, 5

[23] Huggingface. Prompt weighting. https://huggingface.co/docs/diffusers/using-diffusers/weighted_prompts, 2023. 3, 5, 2

[24] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 3

[25] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642, 2023. 3

[26] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. HumanSD: A native skeleton-guided diffusion model for human image generation. In ICCV, 2023. 2

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In CVPR, 2019. 2, 3, 5

[28] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In CVPR, 2023. 3

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In ICLR, 2015. 6

[30] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In CVPR, 2023. 3, 5

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In ICML, 2023. 5

[32] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In CVPR, 2023. 3

[33] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. arXiv preprint arXiv:2310.08579, 2023. 2, 3, 5

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In ICCV, 2015. 2, 3, 5

[35] Samoyed Ventures Pte Ltd. Photo ai. https://photoai.com/, 2023. Accessed: 2023-12-08. 2

[36] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. arXiv preprint arXiv:2307.11410, 2023. 3

[37] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. arXiv preprint arXiv:2303.09319, 2023. 3

[38] Andrew Melnik, Maksim Miasayedzenkau, Dzianis Makarovets, Dzianis Pirshtuk, Eren Akbulut, Dennis Holzmann, Tarek Renusch, Gustav Reichert, and Helge Ritter. Face generation and editing with stylegan: A survey. arXiv preprint arXiv:2212.09102, 2022. 2

[39] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453, 2023. 2, 3

[40] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. TOG, 2022. 2, 3, 5, 6, 1

[41] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In CVPR, 2022. 6

[42] William Peebles and Saining Xie. Scalable diffusion models with transformers. In ICCV, 2023. 3

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023. 2, 3, 5, 1

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In ICML, 2021. 2, 3, 5, 6

[45] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR, 2020. 2, 3

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022. 3

[47] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In EMNLP, 2019. 5

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-

resolution image synthesis with latent diffusion models. In CVPR, 2022. 3

[49] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In CVPR, 2023. 2, 3, 4, 6, 7, 1

[50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. arXiv preprint arXiv:2307.06949, 2023. 2, 3

[51] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, 2022. 2, 3, 4, 5

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In NeurIPS, 2022. 2, 3

[53] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In CVPR, 2015. 6

[54] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021. 3

[55] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402, 2022. 2, 3, 5

[56] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411, 2023. 2, 3

[57] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. In NeurIPS, 2023. 3

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In ICLR, 2021. 3, 6

[59] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In ECCV, 2018. 2

[60] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In ECCV, 2020. 3, 5

[61] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In ICCV, 2023. 3, 4

[62] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431, 2023. 2, 3, 4, 6, 7, 1

[63] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023. 3, 6, 7, 1, 4

[64] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. In NeurIPS, 2023. 3

[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In ICCV, 2023. 2, 3

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018. 6

[67] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In CVPR, 2023. 2

[68] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong

Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In CVPR, 2022. 3, 5

Appendix

Evaluation IDs	
① Alan Turing	⑭ Kamala Harris
② Albert Einstein	⑮ Marilyn Monroe
③ Anne Hathaway	⑯ Mark Zuckerberg
④ Audrey Hepburn	⑰ Michelle Obama
⑤ Barack Obama	⑱ Oprah Winfrey
⑥ Bill Gates	⑲ Renée Zellweger
⑦ Donald Trump	⑳ Scarlett Johansson
⑧ Dwayne Johnson	㉑ Taylor Swift
⑨ Elon Musk	㉒ Thomas Edison
⑩ Fei-Fei Li	㉓ Vladimir Putin
⑪ Geoffrey Hinton	㉔ Woody Allen
⑫ Jeff Bezos	㉕ Yann LeCun
⑬ Joe Biden	

表 3. **用于评估的 ID 名称**。对于每个名称，我们总共收集了四张图像。

A. Dataset Details

训练数据集。根据主文中的 Sec. 3.3，经过一系列过滤步骤，我们构建的数据集中的图像数量约为 112K。这些图像按约 13,000 个 ID 名称进行分类。每张图像都附有对应 ID 的掩码和注释标题。

评估数据集。用于评估的图像数据集包括手动选择的额外 ID 和一部分 MyStyle [40] 数据。对于每个 ID 名称，我们有四张图像作为比较方法的输入数据以及最终指标评估（即，DINO [5]，CLIP-I [14] 和面部相似度 [12]）。对于单嵌入方法（即，FastComposer [62] 和 IPAdapter [63]），我们从每个 ID 组随机选择一张图像作为输入。请注意，评估用的 ID 名称在训练图像集和测试图像集中没有重叠，训练图像集用于我们方法的训练。我们在 Tab. 3中列出了用于评估的 ID 名称。对于用于评估的文本提示，我们考虑了六个因素：服装、配饰、动作、表情、视角和背景，这些因素构成了 40 个提示，具体列在 Tab. 4中。

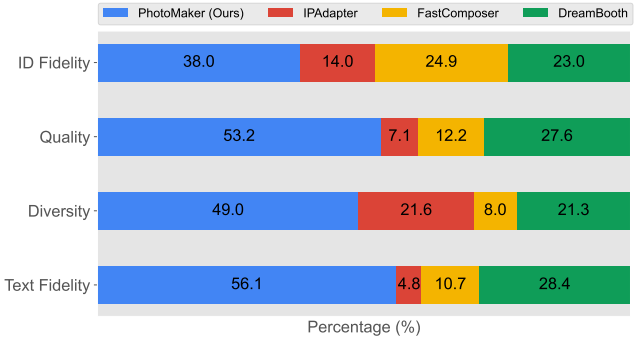


图 9. **不同方法在 ID 保真度、生成质量、面部多样性和文本保真度上的用户偏好**。为便于说明，我们可视化了每种方法所获得的总投票比例。我们的 PhotoMaker 在这四个维度上占据了最大的比例。

B. User Study

在本节中，我们进行了一项用户研究，以便做出更全面的比较。我们选择的比较方法包括 DreamBooth [49]、FastComposer [62] 和 IPAdapter [63]。由于其开源实现，我们使用 SDXL [43] 作为 DreamBooth 和 IPAdapter 的基础模型。我们为每个用户展示了 20 对文本-图像配对，每对配对包括输入 ID 的参考图像和相应的文本提示。对于每个文本-图像配对，我们为每种方法生成了四张随机图像。每个用户需要回答这 20 组结果中的四个问题：1. 哪种方法与输入人物的身份最为相似？2. 哪种方法生成的图像质量最高？3. 哪种方法生成的面部区域最具多样性？4. 哪种方法生成的图像与输入的文本提示最匹配？我们已对所有方法的名称进行了匿名化，并随机排列每组答案中的方法顺序。共有 40 名候选人参与了我们的用户研究，我们收到了 3,200 个有效投票。结果展示在 Fig. 9中。

我们发现，PhotoMaker 在 ID 保真度、生成质量、面部多样性和文本保真度方面具有优势，尤其是在后面三个维度上。此外，我们发现 DreamBooth 在平衡这四个评估维度方面是第二优秀的算法，这也许解释了它为何在过去比基于嵌入的方法更为流行。同时，IPAdapter 在生成图像质量和文本一致性方面存在显著劣势，因为它在训练阶段更侧重于图像嵌入。FastComposer 在面部区域多样性上有明显的不足，因为它采用了单一嵌入训练管道。上述结果与主文中的 Tab. 1大致一致，

Category	Prompt
General	a photo of a <class word>
Clothing	a <class word> wearing a Superman outfit
	a <class word> wearing a spacesuit
	a <class word> wearing a red sweater
	a <class word> wearing a purple wizard outfit
	a <class word> wearing a blue hoodie
Accessory	a <class word> wearing headphones
	a <class word> with red hair
	a <class word> wearing headphones with red hair
	a <class word> wearing a Christmas hat
	a <class word> wearing sunglasses
	a <class word> wearing sunglasses and necklace
	a <class word> wearing a blue cap
	a <class word> wearing a doctoral cap
	a <class word> with white hair, wearing glasses
Action	a <class word> in a helmet and vest riding a motorcycle
	a <class word> holding a bottle of red wine
	a <class word> driving a bus in the desert
	a <class word> playing basketball
	a <class word> playing the violin
	a <class word> piloting a spaceship
	a <class word> riding a horse
	a <class word> coding in front of a computer
	a <class word> playing the guitar

(a)

Category	Prompt
Expression	a <class word> laughing on the lawn
	a <class word> frowning at the camera
	a <class word> happily smiling, looking at the camera
	a <class word> crying disappointedly, with tears flowing
	a <class word> wearing sunglasses
View	a <class word> playing the guitar in the view of left side
	a <class word> holding a bottle of red wine, upper body
	a <class word> wearing sunglasses and necklace, close-up, in the view of right side
	a <class word> riding a horse, in the view of the top
	a <class word> wearing a doctoral cap, upper body, with the left side of the face facing the camera
	a <class word> crying disappointedly, with tears flowing, with left side of the face facing the camera
Background	a <class word> sitting in front of the camera, with a beautiful purple sunset at the beach in the background
	a <class word> swimming in the pool
	a <class word> climbing a mountain
	a <class word> skiing on the snowy mountain
	a <class word> in the snow
	a <class word> in space wearing a spacesuit

(b)

表 4. **按（a）一般设置、服装、配饰、动作，（b）表情、视角和背景分类的评估文本提示**。类别词将被 man、woman、boy 等替换。对于每个 ID 和每个提示，我们随机生成了四张图像用于评估。

除了 CLIP-T 指标的差异。这可能是由于在手动选择与文本中的物体协调的图像时，更倾向于选择与物体协调的图像，而 CLIP-T 则更侧重于物体是否出现。这可能表明了 CLIP-T 的局限性。我们还在 Fig. 14-17中提供了更多的视觉样本供参考。

C. More Ablations

调整身份混合的比例。对于身份混合，我们的方法可以通过控制输入图像池中身份图像的百分比或通过文本加权方法 [18, 23] 来调整合并比例。通过这种方式，我们可以控制生成的新身份与特定输入身份的相似度，既可以使新生成的身份更加接近某一特定输入身份，也可以使其远离该身份。Fig. 10展示了我们通过控制输入图像池中不同身份比例来定制新身份的过程。为了

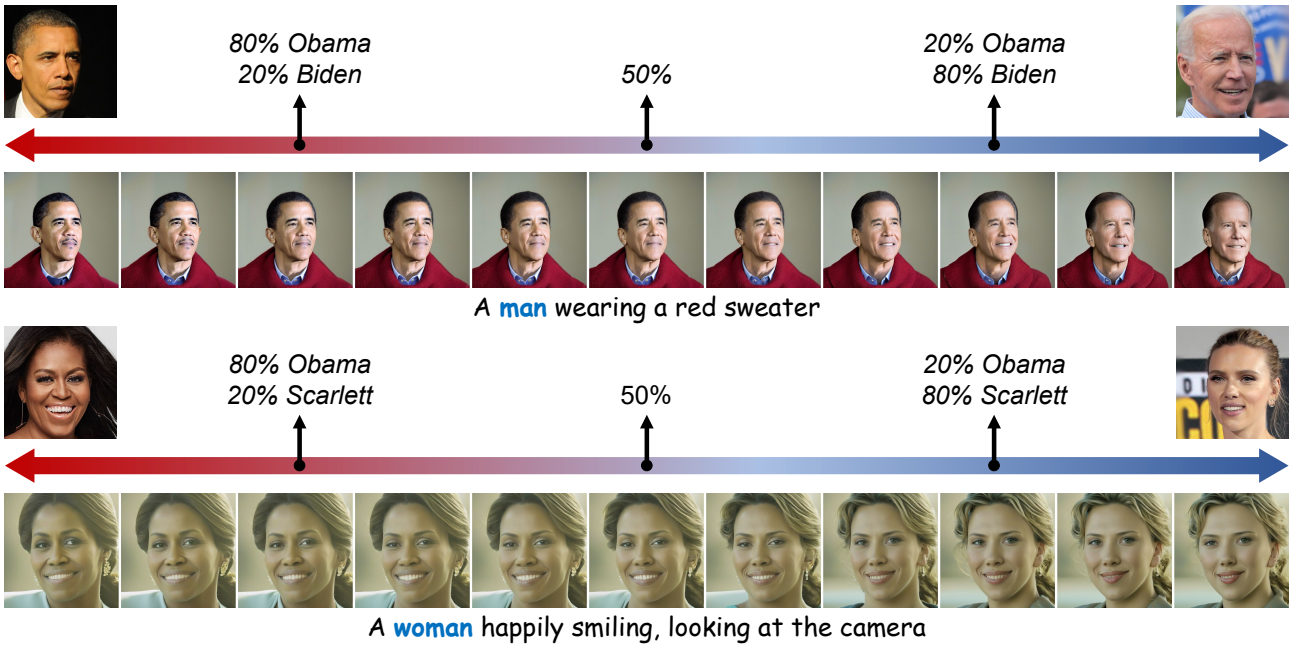


图 10. **不同 ID 图像在输入样本池中所占比例对新 ID 生成的影响**。第一行展示了从 Barack Obama 到 Joe Biden 的过渡。第二行展示了从 Michelle Obama 到 Scarlett Johansson 的变化。为了更清晰地说明，图中使用了百分比来表示每个 ID 在输入图像池中的比例。输入池中包含的总图像数为 10 张。

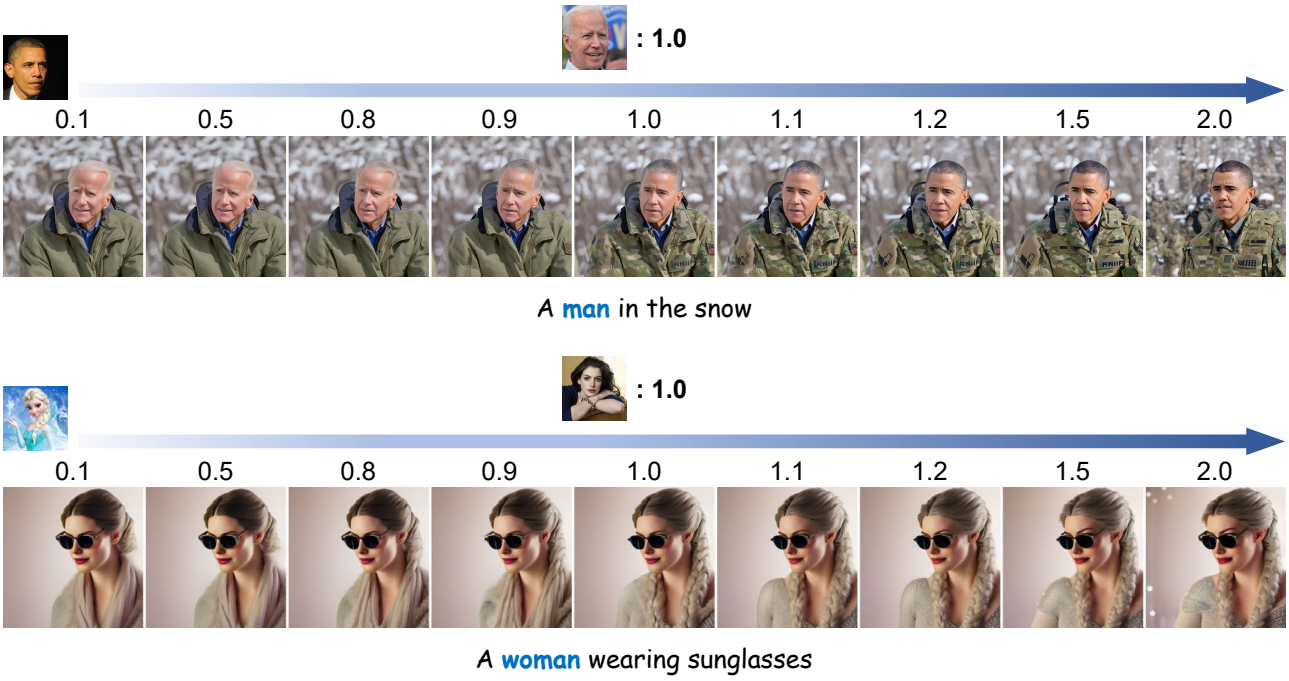


图 11. **文本加权对新 ID 生成的影响**。第一行展示了 Barack Obama 和 Joe Biden 的混合。从左到右的第一行表示逐渐增加对应 Barack Obama 的 ID 图像嵌入的权重。第二行展示了 Elsa（迪士尼）和 Anne Hathaway 的混合。Elsa 的权重逐渐增加。

更好的描述，我们在这个实验中使用了总共 10 张图像作为输入。我们可以观察到在这两个身份之间的平滑过渡。这种平滑过渡包含了皮肤颜色和年龄的变化。接下来，我们使用每个生成的身份四张图像进行文本加权。结果显示在 Fig. 11中。我们通过将与特定身份相关的图像的嵌入向量乘以一个系数，来控制该身份在新身份中的融合比例。与通过控制输入图像数量的方式相比，提示加权需要更少的照片来调整不同身份的合并比例，显示了其优越的可用性。此外，这两种调整不同身份混合比例的方法都展示了我们方法的灵活性。

D. Stylization Results

我们的方法不仅具备生成真实人像的能力，还能够在保持身份属性的同时进行风格化。这展示了所提方法的强大泛化能力。我们在 Fig. 12中提供了风格化的结果。

E. More Visual Results

重新语境化。我们首先在 Fig. 14中提供了一个更直观的比较。我们将我们的 PhotoMaker 与 DreamBooth [49]、FastComposer [62] 和 IPAdapater [63] 进行比较，针对重新语境化的案例。与其他方法相比，我们的方法生成的结果能够同时满足高质量、强文本可控性和高身份保真度的要求。接下来，我们关注 SDXL 无法自行生成的身份。我们将这种场景称为“非名人”案例。在这个设置下，通过比较 Fig. 15和 Fig. 13，我们的方法能够成功生成与输入 ID 一致的图像。

将人物从艺术作品/旧照片带入现实。 Fig. 16-17 展示了我们的方法将过去的名人带回现实的能力。值得注意的是，我们的方法能够从雕像和油画中的 ID 生成逼真的照片。实现这一点对于我们比较的其他方法来说是相当具有挑战性的。

改变年龄或性别。我们在 Fig. 18中提供了更多关于改变年龄或性别的视觉结果。如在正文中所提到的，我们只需在进行此类应用时更改类别词。在改变年龄或性别的生成 ID 图像中，我们的方法能够很好地保留原始 ID 中的特征。

身份混合。我们在 Fig. 19中提供了更多关于身份混合应用的视觉结果。得益于我们堆叠的 ID 嵌入，我们的方法可以有效地将不同 ID 的特征融合，形成一个新的 ID。随后，我们可以基于这个新 ID 生成受文本控制的图像。此外，我们的方法在身份混合过程中提供了极大的灵活性，正如在 Fig. 10-11中所示。更重要的是，我们在正文中探讨了现有方法在实现这一应用时的困难。相反，我们的 PhotoMaker 为这一应用开辟了多种可能性。

F. Limitations

首先，我们的方法只专注于保持图像中单个人物的 ID 信息，无法同时控制图像中多个人物的 ID。其次，我们的方法在生成半身像方面表现出色，但在生成全身像时相对较弱。第三，我们方法的年龄转换能力不如一些基于 GAN 的方法 [1] 精确。如果用户需要更精确的控制，可能需要对训练数据集的描述进行修改。最后，我们的方法基于 SDXL 及我们构建的数据集，因此它也会继承这些模型和数据集的偏差。

G. Broader Impact

在本文中，我们提出了一种新颖的方法，能够生成高质量的真人图像，同时保持与输入身份高度相似的特征。与此同时，我们的方法还具备高效性、良好的面部生成多样性和良好的可控性。

对于学术界而言，我们的方法为个性化生成提供了强有力的基准。我们的数据创建管道可以生成更多样化的数据集，涵盖不同的姿势、动作和背景，这对于开发更强大且具有良好泛化能力的计算机视觉模型具有重要意义。

在实际应用领域，我们的技术有潜力彻底改变娱乐产业，它可以用于为电影或视频游戏创建逼真的角色，而无需大量的 CGI 工作。它在虚拟现实中也具有巨大的应用前景，能够通过让用户在不同场景中看到自己，提供更具沉浸感和个性化的体验。值得注意的是，任何人都可以通过我们的 PhotoMaker 快速定制自己的数字肖像。

然而，我们也认识到生成高保真度人类图像所带来的伦理问题。此类技术的广泛应用可能会导致生成的肖像被不当使用、恶意篡改图像以及传播虚假信息。因此，我们强调开发并遵守伦理准则，负责任地使用此项技术的重要性。我们希望我们的贡献能够激发更多关于计算机视觉中人类生成技术的安全和伦理使用的讨论与研究。

图 12. **我们的** PhotoMaker **方法在不同输入身份和不同文本描述风格下的风格化结果**。我们的方式可以无缝地转移到多种风格，同时防止生成真实的结果。符号 <class> 表示它将根据需要被替换为 man 或 woman。

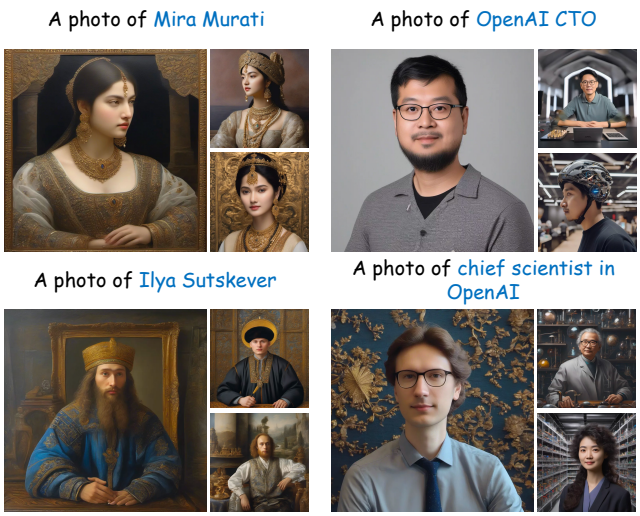


图 13. **两个 SDXL 无法识别的示例**。我们更换了两种文本提示（例如，姓名和职位），但未能成功促使 SDXL 生成 Mira Murati 和 Ilya Sutskever。

图 14. **更多的重新语境化可视化结果**。我们的方法不仅提供了高 ID 保真度，还保留了文本编辑能力。我们为每个文本随机抽取三幅图像。

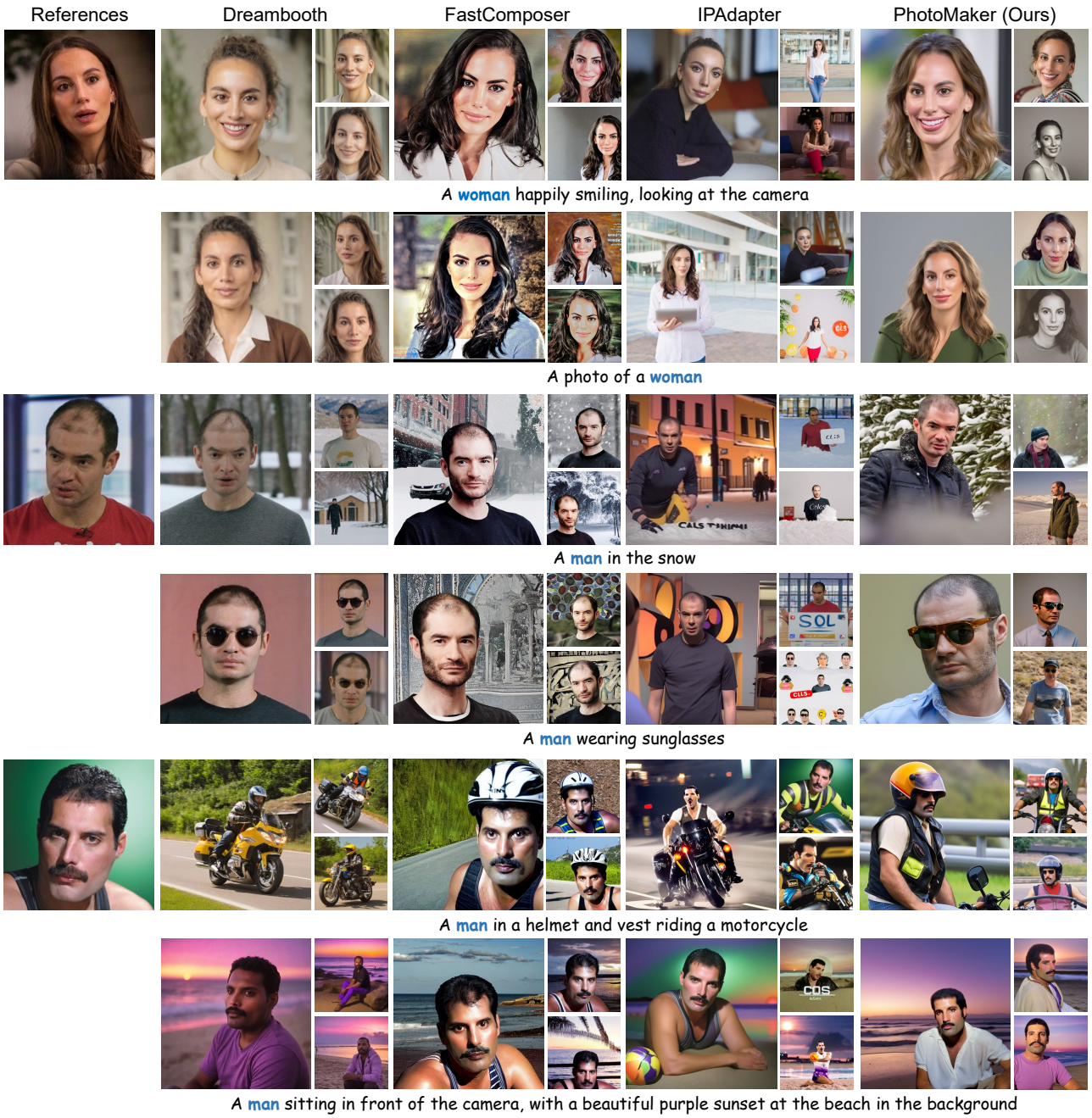


图 15. **更多重新语境化的可视化结果**。我们的方法不仅提供了高 ID 保真度，还保留了文本编辑能力。我们为每个文本随机抽取三幅图像。

图 16. **更多将旧照片中的人物带回现实的可视化结果**。我们的方法可以生成高质量的图像。我们为每个文本随机抽取三幅图像。

图 17. **更多将艺术品中的人物带回现实的可视化结果**。我们的 PhotoMaker 能够生成照片级真实感的图像，而其他方法很难实现这一点。我们为每个文本随机抽取三幅图像。

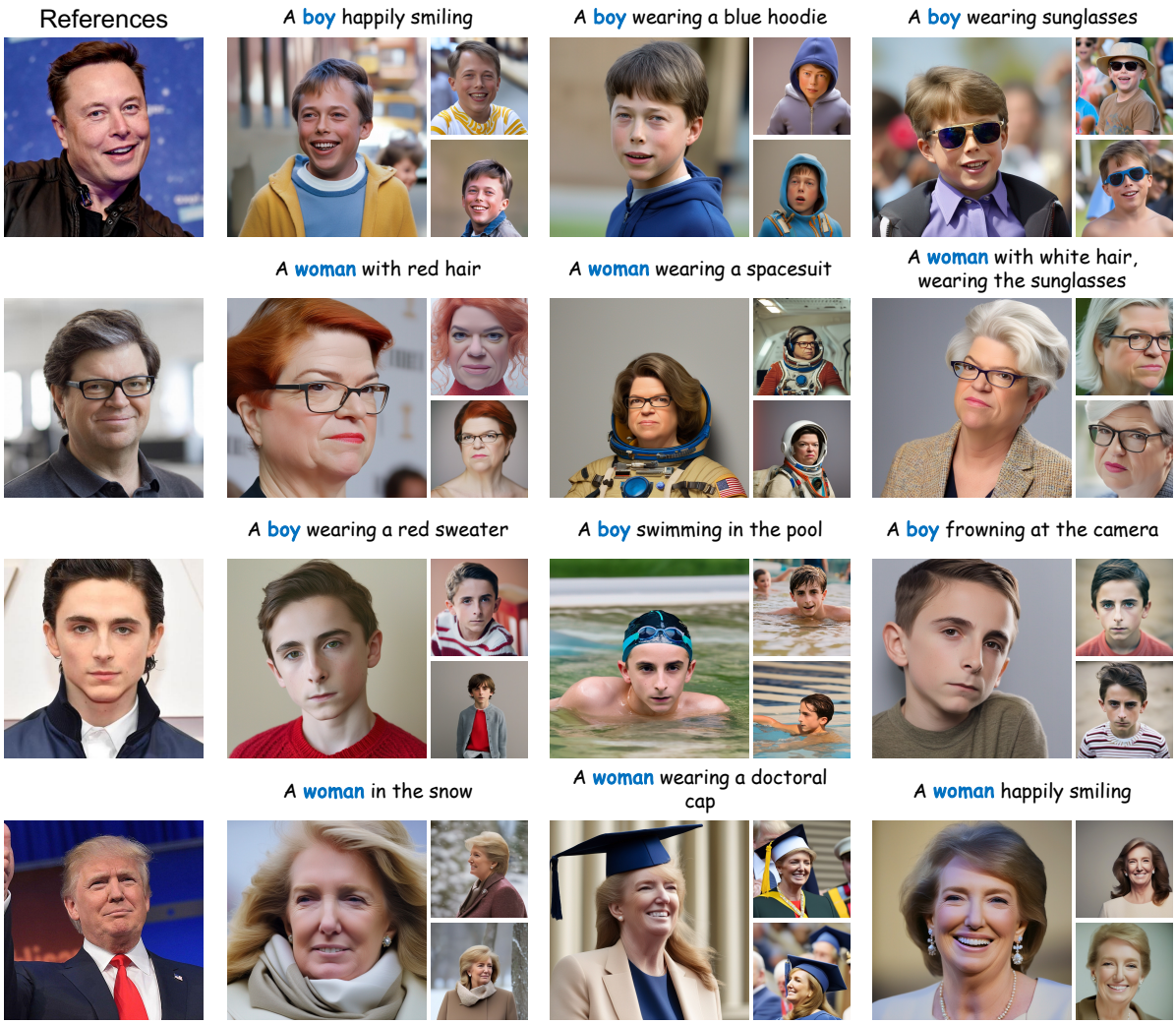


图 18. **更多关于改变每个 ID 的年龄或性别的可视化结果**。我们的 PhotoMaker 在修改输入 ID 的性别和年龄时，能够有效保留面部 ID 的特征，并允许进行文本操作。我们为每个文本随机抽取三幅图像。

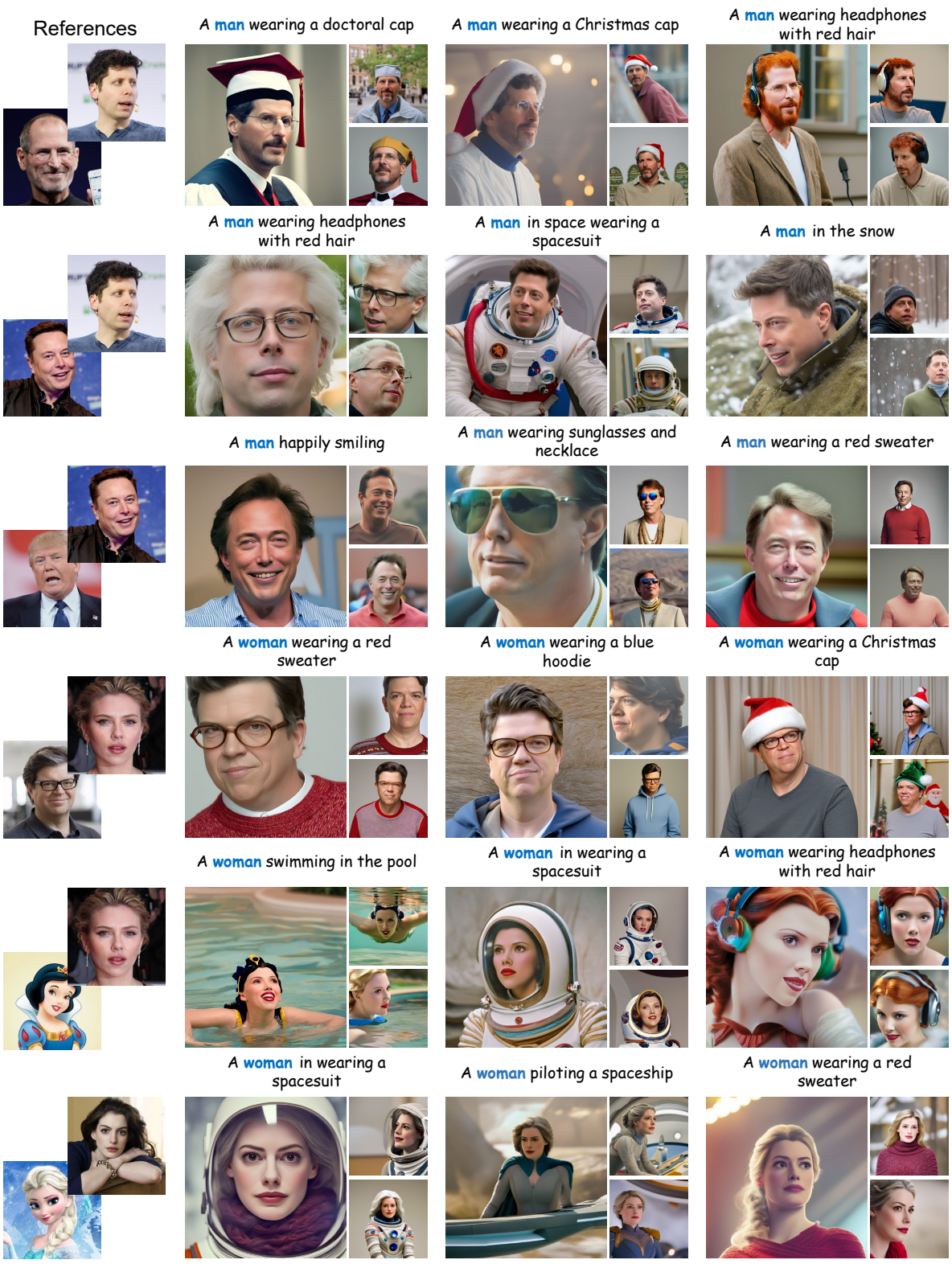


图 19. **更多关于身份混合应用的可视化结果**。 PhotoMaker 能够在新生成的 ID 图像中保持输入 ID 的特征，同时提供高质量且文本匹配的生成结果。我们为每个文本随机抽取三幅图像。